Live Demonstration: Real-Time Spoken Digit Recognition using the DeltaRNN Accelerator

Gao, Chang ; Braun, Stefan ; Kiselev, Ilya ; Anumula, Jithendar ; Delbruck, Tobi ; Liu, Shih-Chii

Abstract: This demonstration shows a real-time continuous speech recognition hardware system using our previously published DeltaRNN accelerator that enables low latency recurrent neural network (RNN) computation. The network is trained on augmented audio samples from the TIDIGITS dataset to achieve a label error rate (LER) of 2.31%. It is implemented on a Xilinx Zynq-7100 FPGA running at 1 MHz. The incremental RNN power consumption is 30 mW. Visitors interact with the system by speaking digits into a microphone connected to the FPGA system and the classification outputs of the network are continuously displayed on a laptop screen in real time.

DOI: https://doi.org/10.1109/iscas.2019.8702212

Accepted for publication at the 2019 IEEE International Symposium on Circuits and Systems (ISCAS)

Live Demonstration: Real-time Spoken Digit Recognition using the DeltaRNN Accelerator

Chang Gao, Stefan Braun, Ilya Kiselev, Jithendar Anumula, Tobi Delbruck and Shih-Chii Liu
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland
Email: [chang, sbraun, kiselev, anumula, tobi, shih]@ini.uzh.ch

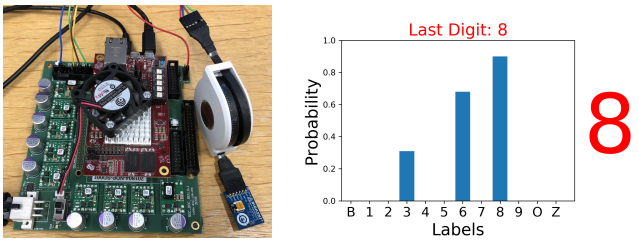
***Abstract*— This demonstration shows a real-time continuous speech recognition hardware system using our previously published DeltaRNN accelerator that enables low latency recurrent neural network (RNN) computation. The network is trained on augmented audio samples from the TIDIGITS dataset to achieve a label error rate (LER) of 2.31%. It is implemented on a Xilinx Zynq-7100 FPGA running at 1 MHz. The incremental RNN power consumption is 30 mW. Visitors interact with the system by speaking digits into a microphone connected to the FPGA system and the classification outputs of the network are continuously displayed on a laptop screen in real time.**

I. Demonstration

This demonstration shows a real-time spoken digit recognition hardware system that runs our previously published DeltaRNN accelerator on a System-on-Chip containing an FPGA [1]. The input from a digital microphone is processed using a frame duration of 25 ms with stride of 10 ms and 15 ms overlap between frames. Frames are normalized by taking the mean and standard deviation of every 8 frames. Features are extracted by applying a 40-dimensional log filter bank on normalized frames. The classification is performed by a delta GRU-RNN [2] with 256 neurons, followed by a 200-dimensional fully-connected (FC) layer with ReLU activation function, which drives an 12-dimensional softmax output layer. Output neurons correspond to the 12 labels (blank, 1–9, Z, O). The network is trained on a dataset based on `TIDIGITS`, augmented by applying speed perturbations and noise injection. The 80h augmented dataset was split into 72h training and 8h test sets. It achieved a label error rate of 2.31% [3].

The delta-GRU layer is computed by the DeltaRNN accelerator running at 1 MHz clock. In low-latency mode [3], the entire spoken digit recognition pipeline latency is 7.47 ms, with only 0.43 ms from the DeltaRNN accelerator. The RNN requires 0.45 MOp per timestep, so the throughput is 1.0 GOp/s (7.7 GOp/s in high-throughput mode [3]). The 0.25 delta threshold reduces weight memory access by 5.2X with negligible loss of accuracy. Audio frame normalization, log filter bank, and FC layers are computed by the ARM core in 7.04 ms. The maximum element of 12-dimensional softmax output vector is sent to a laptop through the USB serial port.

This work was partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 644732, the Swiss National Science Foundation, HEAR-EAR, 200021_172553, and the Samsung Institute of Advanced Technology.



(a) Digit recognition system (b) Classification output.

Fig. 1: Demo system

Total wall power is 11.6 W, but incremental RNN power consumption is only 30 mW at the 1MHz clock frequency, giving incremental power efficiency of 35 GOp/s/W (256.7 GOp/s/W in high-throughput mode).

II. Demonstration Setup

Fig. 1a shows the demo setup. It ues an AVNET Zynq-7100 Mini-Module-Plus with a custom base board[1], with a PmodMIC3 mono-microphone connected through GPIO pins.

III. Visitor Experience

Visitors continuously speak single digits into the microphone and see the classification results of the network on the screen. Fig. 1b shows the results displayed as a bar plot of the most recently classified digits. The most recently recognized digit is shown in two text boxes on the top and right. Videos show the demonstration[2].

References

[1] C. Gao, D. Neil, E. Ceolini, S.-C. Liu, and T. Delbruck, "DeltaRNN: A power-efficient recurrent neural network accelerator," in *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '18. New York, NY, USA: ACM, 2018, pp. 21–30. [Online]. Available: http://doi.acm.org/10.1145/3174243.3174261

[2] D. Neil, J. Lee, T. Delbrück, and S. Liu, "Delta networks for optimized recurrent network computation," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 2584–2593. [Online]. Available: http://proceedings.mlr.press/v70/neil17a.html

[3] C. Gao, S. Braun, I. Kiselev, J. Anumula, T. Delbruck, and S.-C. Liu, "Real-time speech recognition for IoT purpose using a delta recurrent neural network accelerator," *ISCAS 2019 Paper ID: 1274*, 2019.

[1]We gratefully acknowledge the Robotics and Technology of Computers Lab, University of Seville for providing the baseboard

[2]https://www.youtube.com/watch?v=XaNgPUqqDXc and https://youtu.be/5UYOeRxWRWA